Intrinsic Pitch in Opening and Closing Diphthongs of German

Oliver Niebuhr

Institute of Phonetics and Digital Speech Processing
Christian-Albrechts-University, Kiel, Germany
on@ipds.uni-kiel.de

Abstract

Perception experiments using rising and falling F0 slopes in 4 German word-final opening and closing diphthongs show that closing/opening diphthongs support the perception of falling/ rising pitch movements. The latter effect is suggested to be enhanced when the initial close vowel quality is retracted and rounded. Hence, the current knowledge of the relationship between intrinsic pitch and vowel quality seems transferable to diphthongs. The estimated intrinsic pitch values of +- 1.5% are better explained by a psychoacoustic pitch-shift than by a process of speech perception compensating for intrinsic F0.

1. Research Questions

This paper addresses two questions. The first one is, if and in what way intrinsic pitch (as opposed to intrinsic F0) occurs in opening and closing diphthongs of German. If intrinsic pitch was found experimentally, then the second question must be, whether this phenomenon is due to a controlled compensation process of speech perception or the result of a pitch-shift characteristic of the perception of any kind of complex tone.

To the first question, a number of studies from different language backgrounds agree that the pitch of a vowel is not exclusively due to its fundamental frequency, e.g. [1,2,3]. It is also influenced by the location of the vowel in the three dimensional vowel space. The aperture dimension is of major importance. The more open a vowel the higher it is perceived when compared with a close vowel at the same fundamental frequency. The pitch differences are larger when the close vowel is more rounded and/or retracted than the open one. Inasmuch as this contribution to the pitch of a vowel is inherent in the vowel itself, it is referred to as intrinsic pitch.

So far, our knowledge of the existence and the amount of intrinsic pitch is restricted to monophthongs, i.e. subjects had to judge different vowel qualities as part of separate stimuli, interrupted by pauses and/or other segments. It cannot be assumed that our knowledge gained from such a task is simply transferable to diphthongs, where different vowel qualities are to be judged as part of a continuous articulatory and acoustic transition in a holistic unit. Moreover, an acoustic analysis of German closing diphthongs by [4] shows that, in contrast to the onset, stable phases at the offset (if there are any) need not exist for F1 and F2. Further, onset and offset have unequal durations of stable phases and target values in F1 and F2 that differ from those found in monophthongs.

Provided our knowledge of intrinsic pitch in monophhongs *is* applicable to diphthongs, the following hypotheses are tested, using closing and opening diphthongs of German: Since a diphthong mainly consists of a continuous transition from one vowel quality to another, closing diphthongs lead to a gradual decrease of intrinsic pitch. Thus, closing diphthongs support the perception of falling pitch contours. This effect is stronger when the close vowel quality at the offset is retracted and rounded. On the other hand, opening diphthongs lead to a gradual increase of intrinsic pitch and support the perception of rising pitch contours. Analogously, this effect is more pronounced for a back rounded vowel at the onset.

As regards the second question, a high negative correlation was found (cp. [1,3]) between intrinsic pitch and intrinsic F0 values for corresponding monophthongs measured in various languages, e.g. [5,6,7]. Higher/lower intrinsic pitch is connected with lower/higher intrinsic F0. Although the reason for this intrinsic F0 is still controversial, a mechanical linkage between the structures of the larynx and other components of the vocal apparatus involved in the production of vowels (esp. the tongue) is the most probable explanation (cp. [2]). Following this physiological explanation, intrinsic F0 is a highly uncontrollable but consistent phenomenon. Combined with the negative correlation, this led to the hypothesis put forward by [2] that intrinsic pitch is caused by a perceptual process parsing F0 into its intentional and intrinsic component. Since the intrinsic component is used as a cue for segment identification, it is compensated for at the prosodic level.

However, this hypothesis is not supported by the majority of studies dealing with intrinsic pitch. The values found for intrinsic pitch in these studies are much smaller (at most 3% of the F0 value, see [8] for an overview) than those measured for intrinsic F0 under comparable conditions. These intrinsic pitch values are more adequately explained by the general psychoacoustic mechanism of the virtual pitch model [9] valid for any kind of complex tones of which vowels are a subset. In this model, the pitch of a vowel is calculated with reference to its F0 and spectral envelope. Different spectral envelopes, especially F1 but also F2, cause small but perceptible pitch-shifts which are quite similar to the empirical values found for intrinsic pitch.

Nevertheless, some studies, e.g. [10], found values for intrinsic pitch, too large to be explained by pitch-shift alone, thus supporting the hypothesis of compensation. The empirical variations in intrinsic pitch values may be due to the strong variation between subjects reported in every study. Alternatively, the larger intrinsic pitch values may also result from judging (nonsense) words containing the relevant vowels instead of comparing the pitch of synthesized vowels in isolation. The former experiments have a closer relationship to speech and might therefore activate more or different perceptual processes than the latter, more psychoacoustic experiments.

In order to test this hypothesis, resynthesized German words were used in the experiments of this study. Whether the results point to a process of compensation or pitch-shift is a matter of the amount of intrinsic pitch found. However, it is not the primary aim of this study to quantify intrinsic pitch in diphthongs.

2. Method

In order to test the hypotheses of this study, 4 stimulus series were created, each of them with a different German monosyllabic word as its base. In two of the series, the words *"Hai"* ([hai], *'shark'*) and *"hau"* ([hau], *'strike'*) were used, each containing a closing diphthong, ending in a front unrounded or a back rounded vowel quality, respectively. In the remaining two stimulus series the words *"hier"* ([hi:ɐ], *'here'*) and *"Uhr"* ([ʔu:ɐ], *'clock'*) provided the base of the stimuli. Each word contained an opening diphthong, again involving either a front unrounded or a back rounded vowel quality. In these last two diphthongs, the vowel quality [ɐ] is an allophone of /r/ in postvocalic distribution.

Each of the 4 words was chosen out of 30 repetitions, produced in isolation by a male native speaker with an intended flat F0 contour. For the experiments to reveal maximum intrinsic pitch, if found to be due to pitch-shift alone, stimulus generation was guided by two principles: Firstly, the words finally selected from their production row were marked by the greatest movement in F1 and F2. Secondly, the speaker was instructed to produce all flat F0 contours at the lowest possible level. According to [3], this enhances pitch-shift differences.

As a consequence of the production in isolation, the diphthongs of each of the 4 words chosen had a stable offset phase and rather long total duration. To compensate for the latter effect, a linear duration manipulation was carried out, using the psola-resynthesis in *praat* [11]. Thus, the diphthongs were shortened to a total duration more natural for German read speech (cf. [4]). Simultaneously, the 4 diphthongs were approximated to each other in duration. Table 1 summarizes relevant acoustic measurements after this first manipulation for the 4 words.

Table 1: *Acoustic properties of the words used in the 4 stimulus series. 'Dur'= total word duration (ms), 'Diph'= duration of the diphthong (ms), 'F1/F2 on/off'= formant frequencies (rounded to 50Hz) measured 20ms after onset/ before offset of the diphthong, 'Diff F0'= Difference between F0 at F1/F2 onset/offset (Hz).*

	Dur	Diph	F1 on	F1 off	F2 on	F2 off	Diff F0
Hai	340	250	750	250	1100	2000	7
Hau	360	260	650	250	1000	600	10
Hier	350	260	250	600	1900	1050	-8
Uhr	265	265	250	500	600	1000	-12

Table 1 shows that the intrinsic F0 differences measured in the intended flat contours of the 4 words vary between 7Hz and 12Hz. These values are in line with the intrinsic F0 values found in the comparison of individual monophthongs (e.g. [6]). If the compensation hypothesis is valid, the intrinsic pitch will be of a similar (but not much larger) order of magnitude. It was therefore decided to take the *"Uhr"* value of 12Hz as the extremes of an F0 slope continuum. In this continuum, the slope of a linear F0 course varied in 9 equal-sized steps of 3Hz around a constant average of 100Hz. This average value corresponds to the one the words are produced at by the male speaker. For creating the 9 steps, onset and offset of the linear F0 slope were shifted in opposite directions by 1.5Hz each. So, the F0 slope continuum comprises 4 falling and 4 rising F0 courses and a flat one, illustrated in Figure 1.

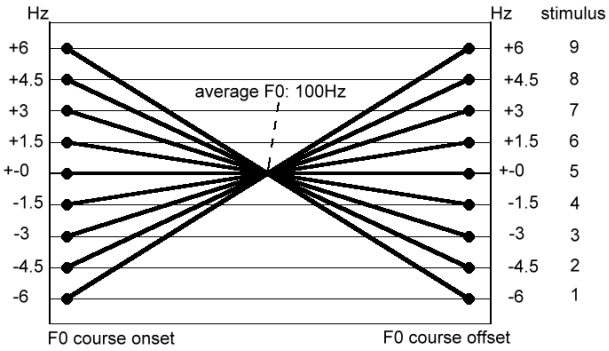

Figure 1: *F0 slope continuum, consisting of 9 steps around a constant average of 100Hz and the corresponding stimulus numbers after applying it to the diphthong of each word.*

By using the psola-resynthesis in *praat* [11], the F0 slope continuum was applied to the diphthongs of the 4 words *"Hai"*, *"hau"*, *"hier"* and *"Uhr"*, resulting in 4 stimulus series, each with 9 stimuli. In all series, stimulus 1 contained an F0 contour, falling linearly over 12Hz, whereas stimulus 9 contained a linear rising F0 movement of the same size. In stimulus 5 the F0 was flat. With reference to [12], the change in slope between adjacent stimuli within a stimulus series is regarded as large enough to be perceived, whereas differences in slope between the same stimulus numbers of different series due to slightly divergent diphthong durations in each word (Tab.1) should be negligible. Furthermore, it had to be considered that the F0 resynthesis in *praat* [11] simply copies voiceless marked parts of the signal under manipulation to the output signal. This also happened for a final part of the diphthong in each word. In order to prevent these unchanged final parts from influencing the judgements of the listeners, they were cut from each stimulus. After that, the decrease in intensity was restored to retain the natural character of the stimuli. This was done by applying the *bell curve* in *cool edit pro* [13] to the last 30ms of the new end of each stimulus.

On the basis of these edited stimuli one ABX-test and one BAX-test was constructed for each stimulus series. In every test, 'A' was stimulus 1 (falling 12Hz) and 'B' was stimulus 9 (rising 12Hz), whereas all stimuli were used for 'X', including 1 and 9. So, in two of the resulting 9 ABX and BAX triads 'X' was physically identical with either 'A' or' 'B'. The elements in each triad were sperated by a 500ms pause. A complete test unit was preceeded by a signal tone, then presented the actual triad twice, separated by 2 seconds (s) pause, and was followed by a 3.5s pause, in which the listeners had to give their judgements. In every test, each unit occurred five times. So, with regard to the whole stimulus series every 'X' was judged 10 times by the listeners, 5 times after 'AB' and 5 times after 'BA'. The 45 units of every test (9 triads x 5 repetitions) were given different randomized orders.

Twelve native speakers of German (8 female, 4 male, average age 26) participated in all 8 experiments (4 stimulus series x 2 tests). They were asked to listen to the tonal movements in the stimuli and judge whether 'X' was more similar to 'A' or 'B'. In this way, the subjects could judge the F0 course in the words as either rising or falling, without explicitly using these potentially misleading terms in the experiments. The paradigm of shifting both ends of the F0 course in opposite directions was intended to support this task, by preventing the subjects from comparing only the endpoints of the contour.

3. Results

The 'X'-stimuli of each stimulus series were judged 10 times by all 12 subjects, resulting in 120 judgements for each 'X'. On the basis of these data, Figure 2 shows the percentage of 'X'-stimuli judged 'more similar to A' at each step of the F0 slope continuum in all 4 stimulus series.

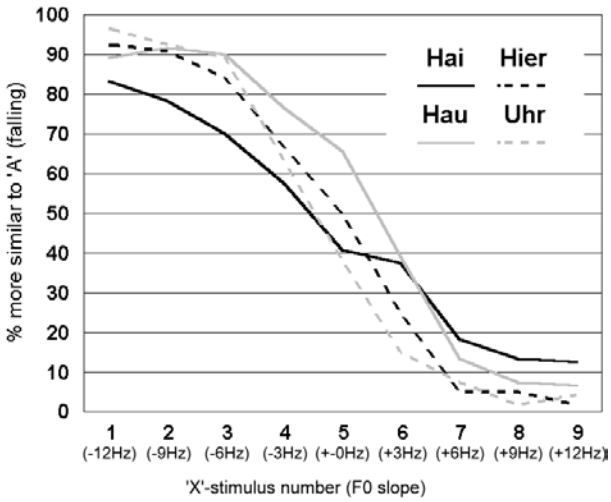

Figure 2:*Percentage of 'X'-stimuli judged more similar to 'A' at each step of the F0 slope continuum of all 4 stimulus series.*

To test the influence of the four diphthongs on the distribution of 'similar to A' or 'similar to B' judgements, respectively, a repeated-measures ANOVA was performed, using two independent variables, each having two possible categories. The first variable was called *DCA* (direction of change in aperture, C=closing, O=opening). The second variable was called *BR* (back/rounded vowel quality, N=not involved, I=involved). Table 2 contains the associations of the diphthongs in *"Hai"*, *"hau"*, *"hier"* and *"Uhr"* with these variables and categories.

To perform the analysis of variance, all 90 judgements of each subject of the 'X'-stimuli in a series (10 judgements x 9 'X'-stimuli) were reduced to only one measurement per subject and series as follows: First, the 'similar to A' responses were arranged by ascending stimulus numbers. Then, going through this arrangement from both ends, the mean (*M*) of those two stimulus numbers was calculated where the value 5 was reached or crossed for the first time. The 48 means created by this procedure (12 subjects x 4 series) constituted the input of the repeated-measures ANOVA. The results of this analysis are summarized in Tables 2 and 3. This procedure is superior to more common ones (e.g. summing up judgements), because it gives a better representation of the expected ogee-like courses of the judgements in each series due to the F0 slope continuum (cf. Fig.2) by extracting the 50%-border between the two response categories.

Table 2: *Descriptive statistics and association of the diphthongs in the 4 words with the two variables and categories used in the repeated-measures ANOVA (see text for abbr.).*

	DCA / BR	**Mean of *M***	**Standard dev.**	**N**
Hai	C / N	4.50	0.929	12
Hau	C / I	5.50	0.603	12
Hier	O / N	4.92	0.597	12
Uhr	O / I	4.54	0.542	12

Table 3: *degrees-of-freedom (df), F statistics and probabilities of α–error (p-values) for the two main effects DCA and BR, their interaction and a contrast.*

	df	**F**	**p**
main effect *DCA* (Hai / Hau) vs. (Hier / Uhr)	1	1.963	<0.189
main effect *BR* (Hai / Hier) vs. (Hau / Uhr)	1	3.541	<0.087
***DCA* * *BR* interaction Hai vs. Hau vs. Hier vs. Uhr**	1	10.160	<0.009**
Contrast Hau vs. (Hier / Uhr)	1	30.670	<0.0001***

4. Discussion

First of all, the results show that identical F0 courses could be judged differently depending in the vowel qualities involved in the underlying diphthong. Provided that the 'more similar to A' judgements based on the perception of falling pitch movements and the 'more similar to B' responses on the perception of rising ones in the 'X'-stimuli, the results allow the following interpretations:

Except for the *"Hai"* series, Figure 2 illustrates that the results of the *"hau"*, *"hier"* and *"Uhr"* series are comparable for the first and the last three stimuli of the slope continuum. For each of these three words, stimuli 1-3 were judged as 'falling' in more than 89% of cases, whereas the stimuli 7-9 were nearly always judged to be 'rising'. In other words, the judgements for the physically identical pairings in a triad (stimuli 1 and 9) do not differ from those of the physically different triads represented by stimuli 2-3 and 7-8, respectively. While these results allow concluding that the influence of different diphthongs on the comparison of pitch movements is negligible for "steep" rising and falling F0 slopes of at least 6Hz (this does not exclude that such an influence can be found by giving the subjects other tasks), the opposite seems true for the flat F0 course and the slow rising and falling F0 slopes of stimuli 4-6. For this part of the F0 slope continuum, the functions of *"hier"* and *"Uhr"* are marked by a faster decrease of 'falling' judgements than the *"hau"* function. For example, the flat F0 course of stimulus 5 received 66% 'falling' judgements in the *"hau"* series (Fig.2), but almost half as much (38%) in the *"Uhr"* series. The other opening diphthong in the *"hier"* series also received a bare majority of 'rising' judgements at stimulus 5 (Fig.2). These differences at stimulus 5 further indicate that the subjects (mostly) relied on pitch movements instead of F0 onsets or offsets for judging 'X' more similar to 'A' or 'B'.

All 12 subjects reported after the experiments that, in the *"Hai"* series, they had problems with comparing 'X' to 'A' and 'B' and so had to guess frequently. This is reflected in the flattened course of the *"Hai"* function in Figure 2 and the larger standard deviation in the repeated-measures ANOVA (Table 2), each compared to the results of the remaining three series. Thus, the results of *"Hai"* are considered as the main reason for finding no significant differences in the repeated-measures ANOVA for the two main effects *DCA* and *BR* but a very significant interaction between them (p<0.009, see Table 3). Consequently, excluding the results for *"Hai"* from the statistical analysis shows for the main effect *DCA* that the point where the judgements change from predominantly

'falling' to 'rising', represented by the means in Table 2, is reached significantly later in the *"hau"* series compared to the *"hier"* and *"Uhr"* series ($p<0.0001$, Table 3, last row). With regard to this finding and because it is problematic to deal with the results of the *"Hai"* series in terms of the hypotheses of this study, it is concluded that the results of this study confirm the hypothesis that the opening diphthongs in *"hier"* and *"Uhr"* support the perception of rising pitch movements, whereas the opposite is true for the closing diphthong in *"hau"*. So, with regard to the aperture dimension, our knowledge from monophthongs is indeed transferable to diphthongs.

Figure 2 further shows that between stimuli 4-6 *"Uhr"* received less 'falling' judgements than *"hier"*. Thus, in an ascending stimulus order the shift from predominantly 'falling' to 'rising' judgements takes place earlier for *"Uhr"* than for *"hier"* (cf. also the means in Table 2). While the deviating formant frequencies in the open vowel qualities [ɐ] (or [a], Table 1) due to the natural production were of minor importance for the former findings, their contribution to this effect must be considered. It can be seen from Table 1 that starting from a comparable F1 in [i] and [u], F1 ends higher for the [ɐ] in *"hier"* than in *"Uhr"*. Hence, if different degrees of opening in the final [ɐ] quality (represented by F1) were responsible for the observed differences between *"hier"* and *"Uhr"*, they should be the other way round. It is therefore likely that the observed effect is mainly caused by the different tongue positions (front vs. back) and labializations (unrounded vs. rounded) in the initial vowel qualities of the two diphthongs (represented by F2). On this basis, the hypothesis that the pitch-shift is enhanced when the close vowel quality is retracted and rounded is suggested to be confirmed by the results. Accordingly, this part of our knowledge of intrinsic pitch in monophthongs also seems to be valid for diphthongs.

Comparing the means of Table 2 with the values of all 4 functions at the 50%-point in Figure 2 reveals that the means mirror these points very well. This firstly shows that the procedure, applied to the total judgements of each of the 12 subjects (see section 3), worked. Secondly, it allows quantifying the intrinsic pitch in all 4 diphthongs on the basis of the means in Table 2. The resulting intrinsic pitch values, presented in Table 4, are of course estimations. However, with regard to the small fluctuations in all formants resulting from the natural production, estimating is more adequate than giving rigid values.

Table 4: *Estimation of the border where responses shift from predominantly 'falling' to 'rising' (a flat pitch was heard) and the corresponding intrinsic pitch for all 4 diphthongs.*

	Diphthong	Border fall/rise	Intr. Pitch
Hai	[ai]	-1.5 Hz	+1.5 %
Hau	[au]	+1.5 Hz	-1.5 %
Hier	[i:ɐ]	-0.25 Hz	+0.25 %
Uhr	[u:ɐ]	-1.38 Hz	+1.38 %

Table 4 shows for example that [u:ɐ] contains a rising intrinsic pitch of about 1.38%. Thus, for perceiving a flat pitch contour F0 should fall about –1.38Hz during the diphthong. The opposite is true for [au]. The estimated values, except those for *"Hai"*, conform with the intrinsic pitch values found by comparing separate synthesized monophthongs (e.g. [3]). Hence, the results for *"hau"*, *"hier"* and *"Uhr"* are regarded as reliable.

Comparing the intrinsic F0 values listed in Table 1 with the values estimated for intrinsic pitch (Table 4) strongly suggests that the latter do not result from a compensation process of speech perception, but rather from a psychoacoustic process of pitch-shift. However, this does not mean that the former compensation process does not exist. Compared to the parsing process assumed by [2], [8] put forward a hypothesis of a more flexible mechanism of F0 processing: The intrinsic F0 information is used either at the segmental or the prosodic level, depending on where it is actually needed. Considering that the stimuli of this study contained rather pronounced formant patterns, whereas the F0 contours were neither suitable for signalling a clear statement nor a clear question or continuation, it is possible that the intrinsic F0 information was used at the prosodic level and compensated for at the segmental level. In that case, inverted stimulus conditions will lead to larger intrinsic pitch values.

5. Outlook

Considering the hypothesis of [8], additional perception experiments should be carried out to test if less pronounced formant patterns in the 4 diphthongs and linear F0 slopes more suitable for the perception of statements and questions will lead to larger intrinsic pitch values. The reason for the deviating results of *"Hai"* must be investigated. Moreover, the results reported in this study should be based on more subjects.

6. References

[1] Chuang, C.-K.; Wang, W.S.-Y., 1978. Psychophysical pitch biases related to vowel quality, intensity difference, and sequential order. *JASA* 64(4), 1004-1014.

[2] Fowler, C.A.; Brown, J.M., 1997. Intrinsic F0 differences in spoken and sung vowels and their perception by listeners. *Perception&Psychophysics* 59(5), 729-738.

[3] Stoll, G., 1984. Pitch of vowels: experimental and theoretical investigation of its dependence on vowel quality. *Speech Communication* 3, 137-150.

[4] Simpson, A.P., 1998. Phonetische Datenbanken des Deutschen in der empirischen Sprachforschung und der phonologischen Theoriebildung. *AIPUK* 33.

[5] Ladd, D.R.; Silverman, K.E.A., 1984. Vowel intrinsic pitch in connected speech. *Phonetica* 41, 31-40.

[6] Antoniadis, Z.; Strube, H.W., 1981. Untersuchungen zum 'intrinsic pitch' deutscher Vokale. *Phonetica* 38, 277-290.

[7] Lehiste, I.; Peterson, G.E., 1961. Some basic considerations in the analysis of intonation. *JASA* 34(4), 419-425.

[8] Reinholt Petersen, N., 1986. Perceptual compensations for segmentally conditioned fundamental frequency perturbations. *Phonetica* 43, 31-42.

[9] Terhardt, E., 1973. Pitch, consonance and harmony. *JASA* 55(5), 1061-1069.

[10] Rosenvold, E., 1981. The role of intrinsic F0 and duration in the perception of stress. *ARIPUC* 15, 147-166.

[11] Boersma, P. and D. Weenink. Praat: doing phonetics by computer. http://www.fon.hum.uva.nl/praat/

[12] Klatt, D.H., 1973. Discrimination of fundamental frequency contours in synthetic speech: implications for models of pitch perception. *JASA* 53, 8-16.

[13] For more information, see www.cooledit.com